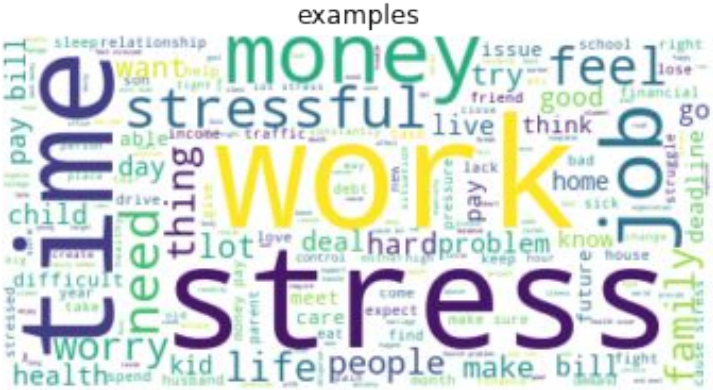
Data Exploration

word cloud

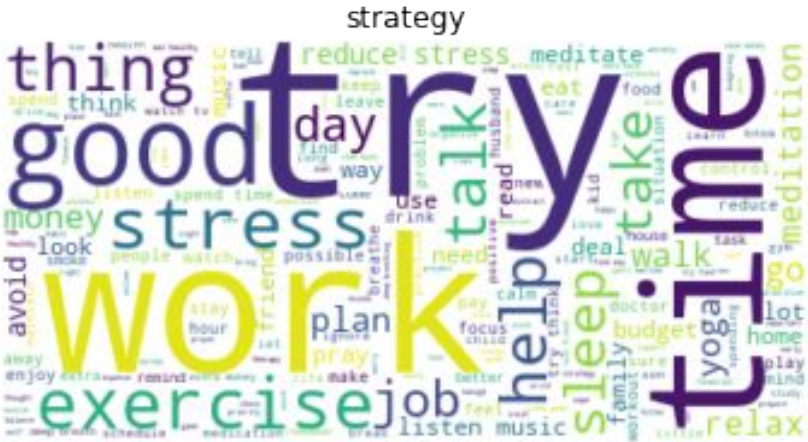
stresses

examples



Data Exploration

word cloud



Data Exploration

LDA (Latent Dirichlet Allocation) for Topic Modeling on Examples Data

```
[(0,
  '0.023*"crazy" + 0.022*"ignore" + 0.021*"wall" + 0.020*"brace" + 0.018*"struggle" + 0.018*"family" + 0.016*"soon" + 0.015*"people" + 0.015*"stong"
+ 0.015*"built"'),
 (1,
  '0.027*"earn" + 0.023*"uncertain" + 0.023*"expectation" + 0.022*"need" + 0.020*"family" + 0.019*"society" + 0.019*"late" + 0.016*"live" +
0.015*"meet" + 0.014*"raise"'),
 (2,
  '0.059*"not" + 0.035*"time" + 0.030*"like" + 0.023*"job" + 0.022*"work" + 0.017*"stress" + 0.016*"goal" + 0.014*"bill" + 0.014*"money" +
0.014*"individually"'),
 (3,
  '0.043*"friend" + 0.025*"bill" + 0.022*"worry" + 0.019*"able" + 0.018*"unreasonable" + 0.018*"job" + 0.018*"diagnose" + 0.016*"pay" + 0.014*"go" +
0.014*"stage"'),
 (4,
  '0.036*"sick" + 0.014*"issue" + 0.014*"depress" + 0.014*"uncontrolled" + 0.013*"cornered" + 0.013*"lead" + 0.013*"feel" + 0.013*"son" +
0.012*"health" + 0.012*"please"'),
 (5,
  '0.022*"responsibility" + 0.017*"worry" + 0.016*"income" + 0.016*"irritate" + 0.016*"education" + 0.016*"wedding" + 0.016*"swell" +
0.015*"condition" + 0.015*"financial" + 0.015*"prepare"'),
 (6,
  '0.020*"n" + 0.019*"feel" + 0.018*"health" + 0.016*"ask" + 0.014*"recently" + 0.014*"organize" + 0.014*"trap" + 0.014*"boy" + 0.014*"divorce" +
0.013*"worry"'),
 (7,
  '0.022*"anger" + 0.020*"manage" + 0.019*"recent" + 0.018*"time" + 0.018*"date" + 0.017*"plan" + 0.016*"specific" + 0.016*"cancer" +
0.016*"project" + 0.015*"try"')]
```

Data Issues

- Spelling issues: ciggarette, personnaly, noice, costumer(?), tak(?), startegie
- Arguably unhealthy stress reduction strategies:
  - i take drugs
  - I fo to the doctor and take medications.
  - I smoke to relax
  - Doing drugs could get your mind off things.
  - i use a lot of alcohol
  - I usually have a few alcoholic beverages then I don't care as much
  - i spend time with my friends to drink and smoke at night.

- Phrases that might not generalize for others
  - stress: No money , example: i don't have a job, strategy: There is no time for me ever.
  - example: My own job is to get my family problems, strategy: finance.
  - example: Finances, strategy: Breathing

Approach

Word Embeddings GloVe

- How it works
  - each word represented as a vector
  - for a phrase, take the average of the vector values
  - similar phrases will have similar average vectors
  - similarity measures, like cosine similarity, show how similar the phrases are
- Example

```
doc1 = nlp('I want to go to the gym')
doc2 = nlp("I only go to the gym when I am free")
doc3 = nlp('This is a zebra at the zoo')

print(doc1.similarity(doc2))
print(doc1.similarity(doc3))
```

```
0.9389000618057138
0.7155503451549461
```

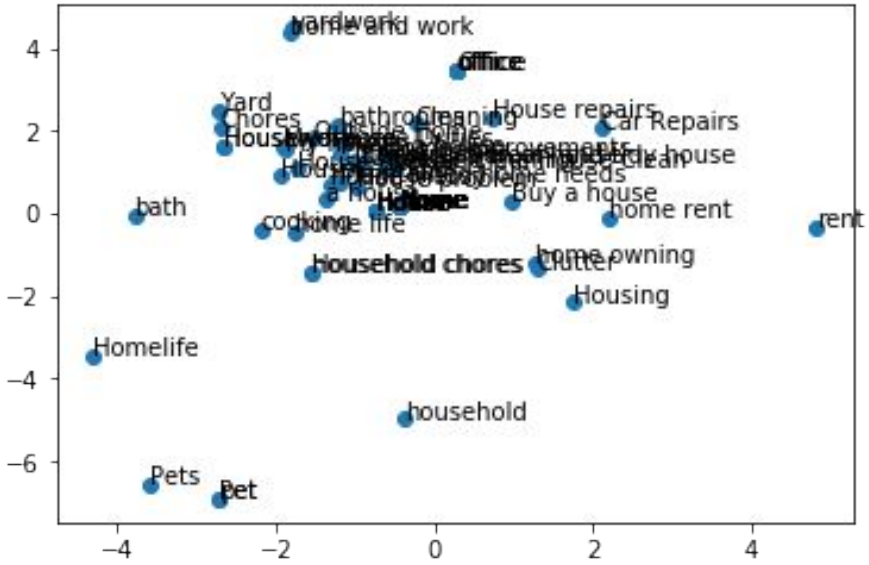
In [493]:

Cluster Phrases with KMeans

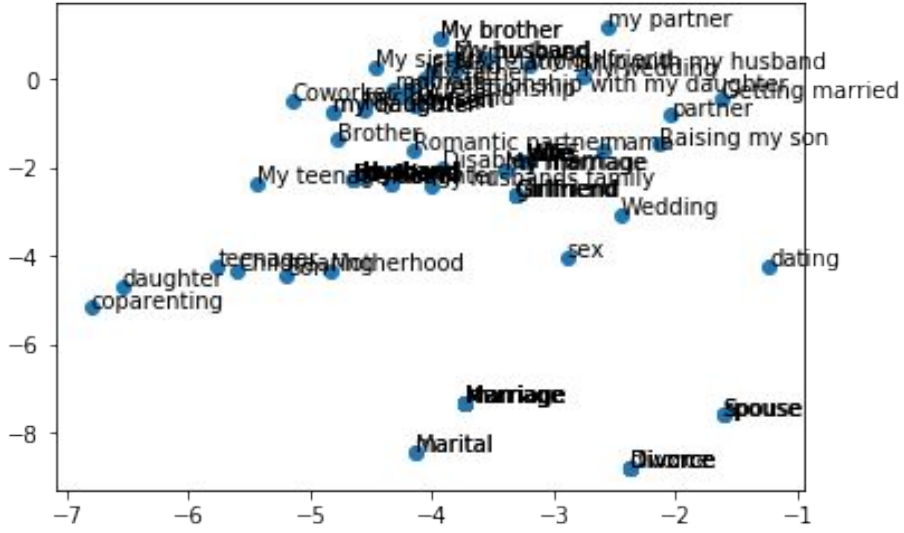
Word Embeddings GloVe

Could use this to trim/fix the data set and/or focus on answers to these topics

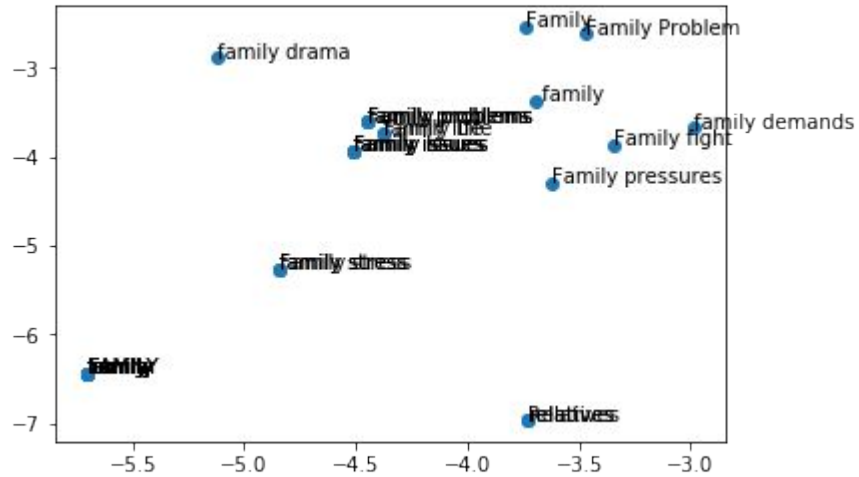
Chores and home responsibilities



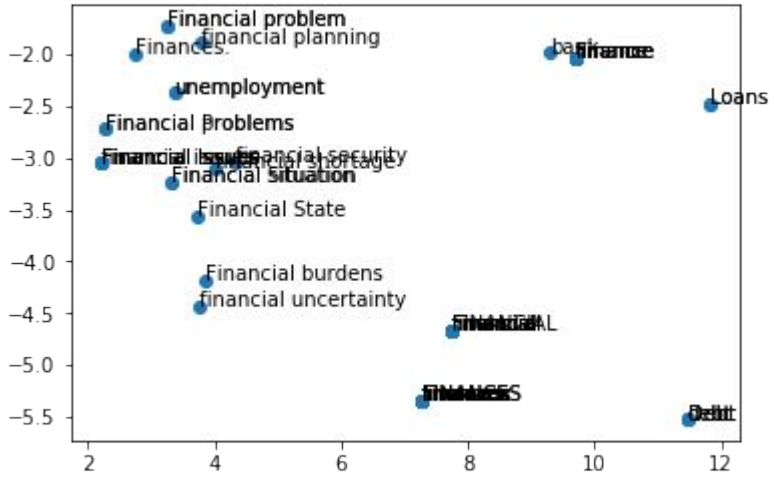
Personal relationships



Family stress



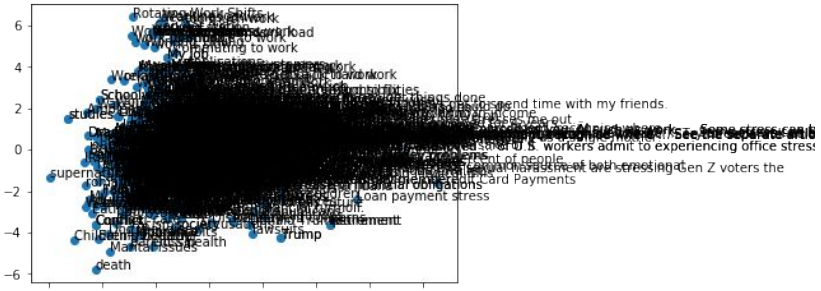
Finance



Other Clusters Include:

- Transportation (Commuting, Parking, Subway)
- Education, School, College, University
- Poor health, mental health
- Childcare and children
- Romance and friendships
- Mental health diseases
- Work and job issues
- Eating and Drinking
- Time management
- Fear of uncertain future

Too many values → cluster only these values



Model Approach

Word Embeddings GloVe

1. With more time, I'd look at clusters of stressors and examples, clean the data, pick out the appropriate strategies, and use a chatbot API like IBM Watson
2. The MVP version does the following:
    a. Reads in a new stressor/example
    b. Calculates the GloVe embedding
    c. Finds the stressor or examples that match the closest
    d. Outputs the top 5 strategies

Example Results